# Unit-3
# Data Preprocessing

## Introduction

***Data preprocessing*** is a data mining technique that involves transforming raw data into an understandable format.

Raw data (real-world data) is often incomplete, inconsistent, and/or noisy, due to which there are some increased chances of error and misinterpretation

- ***Incomplete:*** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. E.g., occupation = " "
- ***Noisy:*** containing errors or outliers. E.g. Salary = "-10"
- ***Inconsistent:*** containing discrepancies in codes or names. E.g. Age="42" Birthday="03/07/1997"

***Data preprocessing*** is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

## *Why is data dirty?*

Raw data collected from the environment can be dirty. A data is said to be dirty if it is incomplete, noisy, or inconsistent.

- ***Incomplete*** data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- ***Noisy data*** (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- ***Inconsistent*** data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

## *Why do we need to preprocess data?*

By preprocessing data, we:

- ***Make our database more accurate:*** We eliminate the incorrect or missing values that are there as a result of human factor or bugs.
- ***Boost Consistency:*** when there are inconsistencies in data or duplicates, it affects the accuracy of the results.
- ***Make the database more complete:*** We can fill the attributes that are missing if needed.
- ***Smooth the data:*** This way we make it easier to use and interpret.

## Major Tasks in Data Preprocessing

| | |
|---|---|
| ***1. Data Cleaning*** | It can be applied to remove noise and correct inconsistences in the data. |
| ***2. Data Integration*** | It combines data from multiple sources into a coherent data store. E.g. data warehouse. |
| ***3. Data Transformation*** | Data Transformation is the process of consolidation of data so that the mining process result could be applied or maybe more efficient. |
| ***4. Data Reduction*** | Obtains reduced representation in volume but produces the same or similar analytical results. |
| ***5. Data Discretization*** | Part of data reduction but with particular importance, especially for numerical data. |

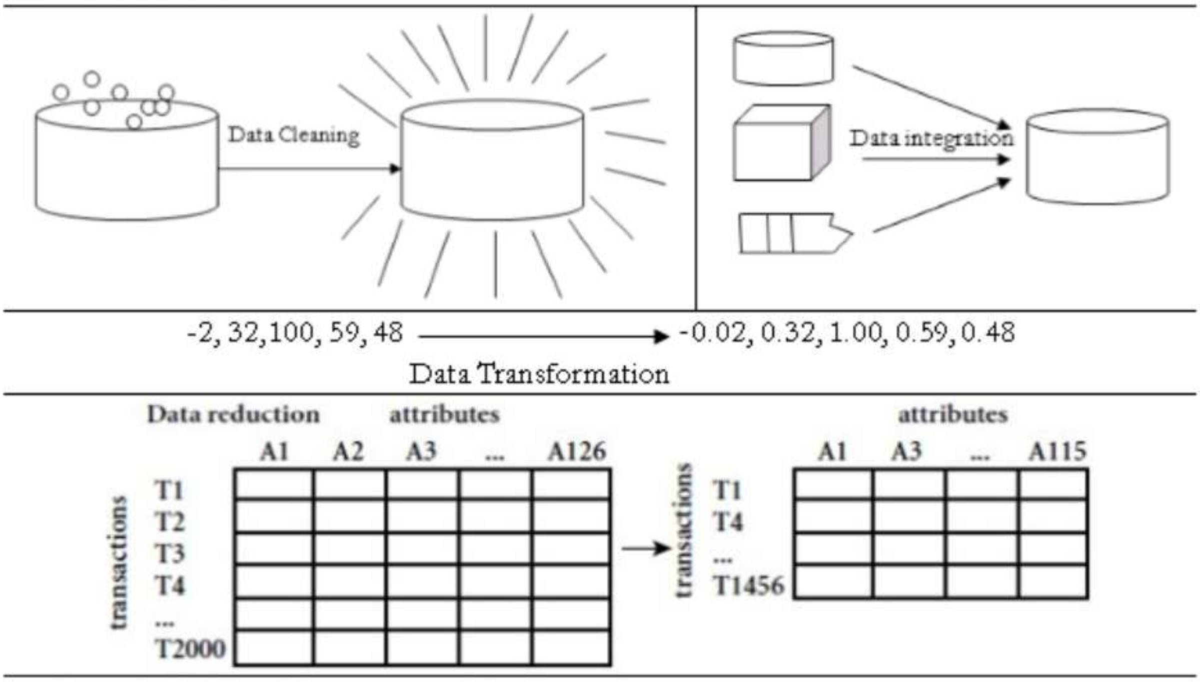



***Fig: Forms of data preprocessing***

## Data Cleaning

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

**a)** ***Missing Data:*** This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

- Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
- Fill in the missing value manually.
- Use a global constant to fill in the missing value. E.g. "unknown", a new class.
- Use the attribute mean to fill in the missing value.
- Use the attribute mean for all samples belonging to the same class as the given tuple.
- Use the most probable value to fill in the missing value.

b) ***Noisy Data:*** Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. E.g. Salary = "-10". It can be handled in following ways:

- ***Binning method:*** This method is to smooth or handle noisy data. First, the data is sorted then and then the sorted values are separated into segments of equal size and stored in the form of bins. There are three methods for smoothing data in the bin.
  - ***Smoothing by bin mean method:*** In this method, the values in the bin are replaced by the mean value of the bin;
  - ***Smoothing by bin median:*** In this method, the values in the bin are replaced by the median value;
  - ***Smoothing by bin boundary:*** In this method, the using minimum and maximum values of the bin values are taken and the values are replaced by the closest boundary value.

  ***Example:***

  Unsorted data for price in dollars: 8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34
  After Sorting: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

  Partitioning into (equal-frequency) bins:
  Bin 1: 8, 9, 15, 16
  Bin 2: 21, 21, 24, 26,
  Bin 3: 27, 30, 30, 34

| ***Smoothing by bin means*** | ***Smoothing by bin boundaries*** |
|---|---|
| Bin 1: 12, 12, 12, 12 | Bin 1: 8, 8, 16, 16 |
| Bin 2: 23, 23, 23, 23 | Bin 2: 21, 21, 26, 26, |
| Bin 3: 30, 30, 30, 30 | Bin 3: 27, 27, 27, 34 |

- ***Regression:*** Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
- ***Clustering:*** This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

## Data Integration

***Data Integration*** is a data preprocessing technique that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data. These sources may include multiple data cubes, databases, or flat files.

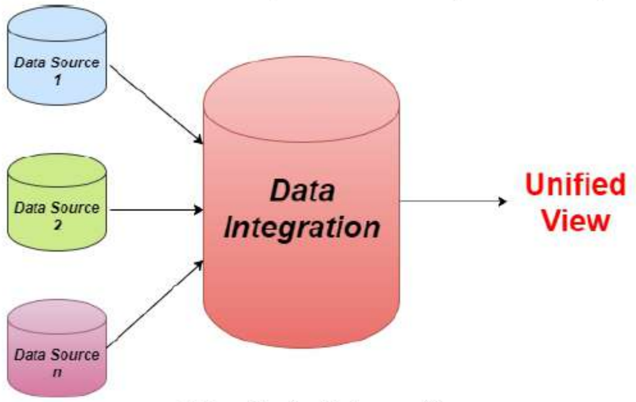



***Fig: Data Integration***

There are mainly ***2 major approaches*** for data integration:

- ***Tight Coupling:*** In tight coupling data is combined from different sources into a single physical location through the process of ETL - Extraction, Transformation and Loading.
- ***Loose Coupling:*** In loose coupling data only remains in the actual source databases. In this approach, an interface is provided that takes query from user and transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.

***Issues in Data Integration***

- ***Entity Identification Problem:*** As we know the data is unified from the heterogeneous sources then how can we 'match the real-world entities from the data'? For example, we have customer data from two different data source. An entity from one data source has *customer_id* and the entity from the other data source has *customer_number*. Now how does the data analyst or the system would understand that these two entities refer to the same attribute?
- ***Redundancy:*** An attribute may be redundant if it can be derived or obtaining from another attribute or set of attributes. Inconsistencies in attributes can also cause redundancies in the resulting data set. Some redundancies can be detected by correlation analysis.
- ***Data Conflict Detection and Resolution:*** Data conflict means the data merged from the different sources do not match. Like the attribute values may differ in different data sets. The difference maybe because they are represented differently in the different data sets. For suppose the price of a hotel room may be represented in different currencies in different cities. This kind of issues is detected and resolved during data integration.

## Data Transformation

Data transformation is the process of transforming data into the form that is appropriate for mining.

***Some Data Transformation Strategies***

1. ***Smoothing:*** It is used to remove the noise from data. Such techniques include binning, clustering, and regression.
2. ***Aggregation:*** Here summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
3. ***Generalization:*** Here low level data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher level concepts, like city or country.
4. ***Attribute construction:*** Here new attributes are constructed and added from the given set of attributes to help the mining process.
5. ***Normalization:*** Here the attribute data are scaled so as to fall within a small specified range, such as -1 to +1, or 0 to 1. Techniques that are used for normalization are:

- ***Min-Max Normalization:*** It performs a linear transformation on the original data. Suppose that min and max are the minimum and maximum values of an attribute, $A$. Min-max normalization maps a value, $v$ , of $A$ to $nv$ in the range [*new_min, new_max*] using following formula.

$$nv = \frac{v - \min}{\max - \min}(new_\max - new_\min) + new_\min$$

- ***Z-score Normalization:*** In z-score normalization (or zero-mean normalization), the values for an attribute, $A$ , are normalized based on the mean and standard deviation of $A$. The value, $v$ , of $A$ is normalized to $nv$ as below. It is also called standard normalization.

$$nv = \frac{v - \mu}{\sigma} \quad \text{Where,} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \mu)^2}{N}}.$$

where, $\mu$ is mean and $n$ is number of data points.

***Example***

Marks: 8, 10, 15, 20

***Min-Max Normalization:***

$Min = 8, Max = 20, v$ is the respective value of the attribute.
$new_max = 1, new_min = 0$

For marks as 8:
$nmark_8 = \frac{8-8}{20-8}(1-0) + 0 = 0$

For marks as 10:
$nmark_{10} = \frac{10-8}{20-8}(1-0) + 0 = 0.16$

For marks as 15:
$nmark_{15} = \frac{15-8}{20-8}(1-0) + 0 = 0.58$

For marks as 20:
$nmark_{20} = \frac{20-8}{20-8}(1-0) + 0 = 1$

Marks after min-max normalization: 0, 0.16, 0.58, 1

***Z-score Normalization:***

$\mu = (8 + 10 + 15 + 20)/4 = 13.25$

$$\sigma = \sqrt{\frac{(8-13.25)^2 + (10-13.25)^2 + (15-13.25)^2 + (20-13.25)^2}{4}} = 4.6$$

Now,

$nmark_8 = \frac{8-13.25}{4.6} = -1.14$ $\quad$ $nmark_{10} = \frac{10-13.25}{4.6} = -0.7$

$nmark_{15} = \frac{15-13.25}{4.6} = 0.3$ $\quad$ $nmark_{20} = \frac{20-13.25}{4.6} = 1.4$

Marks after z-score normalization: $-1.14$, $-0.7$, 0.3, 1.4

## Data Reduction

A database or date warehouse may store terabytes of data. So it may take very long to perform data analysis and mining on such huge amounts of data. Data Reduction is obtaining a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results.

### *Data Reduction Techniques*

1. ***Dimensionality Reduction:*** Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration. Dimensionality reduction methods include *wavelet transforms* and *principal components analysis*, which transform or project the original data onto a smaller space. *Attribute subset selection* is a method of dimensionality reduction in which irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed. For example,

| Name | Mobile No. | Mobile Network |
|---|---|---|
| Jayanta | 9843xxxxxx | NTC |
| Sagar | 9801xxxxxx | NCELL |

*Fig: Before Dimension Reduction*

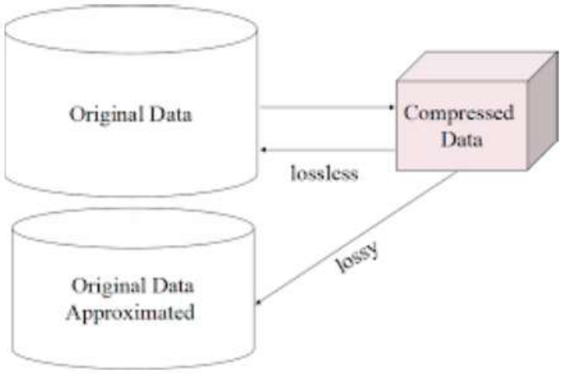
If we know Mobile Number, then we can know the Mobile Network. So we need to reduce the one dimension.

| Name | Mobile No. |
|---|---|
| Jayanta | 9843xxxxxx |
| Sagar | 9801xxxxxx |

*Fig: After Dimension Reduction*

2. ***Numerosity Reduction:*** Numerosity reduction techniques replace the original data volume by alternative, smaller forms of data representation. These techniques may be *parametric* or *nonparametric*. For ***parametric methods***, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data (Outliers may also be stored.) Regression and log-linear models are examples. ***Nonparametric methods*** for storing reduced representations of the data include histograms, clustering, sampling, and data cube aggregation.

3. ***Data Compression:*** In data compression, transformations are applied so as to obtain a reduced or "compressed" representation of the original data. If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called ***lossless***. If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called ***lossy***.



## Data Discretization and Concept Hierarchy Generation

### Data Discretization

***Discretization*** reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

***Example:***

Suppose we have an attribute of Age with the given values:

| | |
|---|---|
| Age | 1, 4, 5, 7, 9, 11, 14, 17, 18, 19, 21, 31, 33, 36, 42, 44, 46, 70, 74, 77, 78 |

***Before Discretization***

| Attribute | Age | Age | Age | Age |
|---|---|---|---|---|
| | 1, 4, 5, 7, 9 | 11, 14, 17, 18, 19, 21 | 31, 33, 36, 42, 44, 46 | 70, 74, 77, 78 |
| ***After Discretization*** | Child | Young | Mature | Old |

Discretization can be categorized into following two types:

- ***Top-down discretization:*** If we first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat of this method up to the end, then the process is known as top-down discretization also known as splitting.
- ***Bottom-up discretization:*** If we first consider all the constant values as split-points, some are discarded through a combination of the neighborhood values in the interval, that process is called bottom-up discretization.

### Concept Hierarchies

***Concept Hierarchies*** reduce the data by collecting and replacing low level concepts (such as city) by higher level concepts (such as province or country).

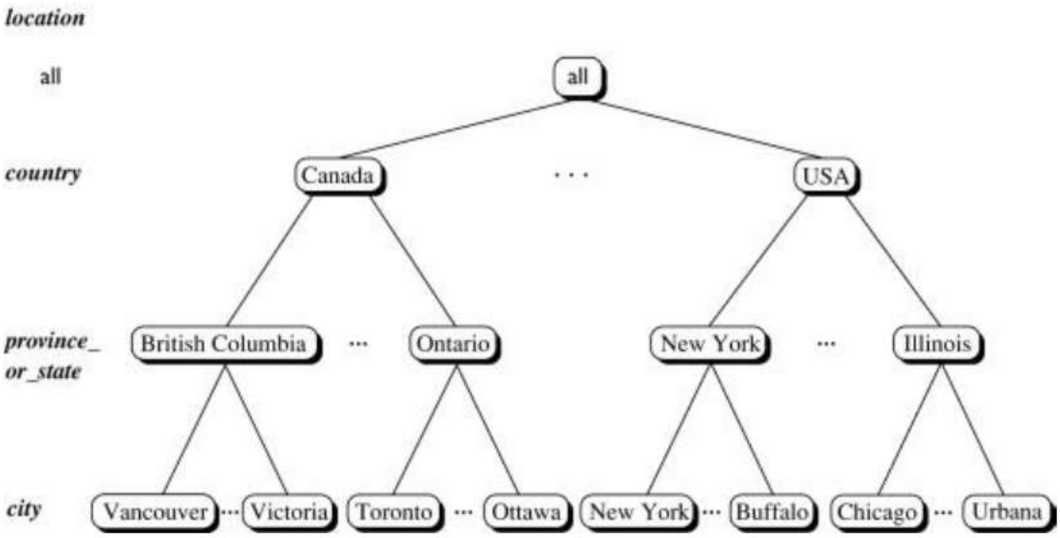



***Fig: Concept Hierarchy***

## *Techniques of Discretization and Concept Hierarchy Generation for Numerical Data*

- ***Binning:*** Attribute values can be discretized by distributing the values into bin and replacing each bin by the mean bin value or bin median value. These technique can be applied recursively to the resulting partitions in order to generate concept hierarchies.
- ***Histogram Analysis:*** Histograms can also be used for discretization. Partitioning rules can be applied to define range of values. The histogram analysis algorithm can be applied recursively to each partition in order to automatically generate a multilevel concept hierarchy, with the procedure terminating once a pre-specified number of concept levels has been reached.
- ***Clustering:*** A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a node of a concept hierarchy, where all noses are at the same conceptual level. Each cluster may be further decomposed into sub-clusters, forming a lower level in the hierarchy. Clusters may also be grouped together to form a higher-level concept hierarchy.

## Data Mining Task Primitives

We can specify a data mining task in the form of a data mining query. This query is input to the system.
A data mining query is defined in terms of data mining task primitives. These primitives allow us to communicate in an interactive manner with the data mining system.

The data mining task primitives are:

1. ***Task-relevant data:*** This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).
2. ***The Kind of knowledge to be mined:*** This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.
3. ***Background Knowledge:*** Background knowledge is information about the domain to be mined that can be useful in the discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.
4. ***Interestingness measures:*** These functions are used to separate uninteresting patterns from knowledge. They may be used to guide the mining process, or after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures.
5. ***Presentation and visualization of discovered patterns:*** This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

Please let me know if I missed anything or anything is incorrect.

poudeljayanta99@gmail.com